

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 2 年 1 2 月 1 2 日
Date of Application:

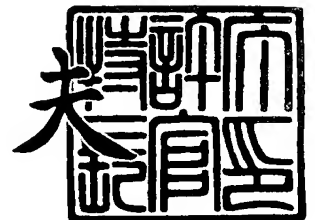
出 願 番 号 特 願 2 0 0 2 - 3 6 0 9 8 4
Application Number:
[ST. 10/C] : [J P 2 0 0 2 - 3 6 0 9 8 4]

出 願 人 セイコーエプソン株式会社
Applicant(s):

2 0 0 3 年 1 0 月 1 0 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 3 - 3 0 8 3 9 1 7

【書類名】 特許願

【整理番号】 J0095689

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/00

【発明者】

 【住所又は居所】 長野県諏訪市大和 3 丁目 3 番 5 号 セイコーエプソン株式会社内

 【氏名】 萱原 直樹

【発明者】

 【住所又は居所】 長野県諏訪市大和 3 丁目 3 番 5 号 セイコーエプソン株式会社内

 【氏名】 大橋 洋貴

【特許出願人】

 【識別番号】 000002369

 【氏名又は名称】 セイコーエプソン株式会社

【代理人】

 【識別番号】 100095728

 【弁理士】

 【氏名又は名称】 上柳 雅誉

 【連絡先】 0 2 6 6 - 5 2 - 3 1 3 9

【選任した代理人】

 【識別番号】 100107076

 【弁理士】

 【氏名又は名称】 藤網 英吉

【選任した代理人】

 【識別番号】 100107261

 【弁理士】

 【氏名又は名称】 須澤 修

【手数料の表示】

【予納台帳番号】 013044

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 0109826

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書抽出装置及び文書抽出プログラム並びに文書抽出方法

【特許請求の範囲】

【請求項 1】 抽出候補となる複数の文書を取得し、それら各文書間の全ての類似度を算出する類似度計算手段と、その文書群のなかから任意の数の文書を抽出するに際し、この類似度計算手段で算出された各文書間の類似度の総和が最も小さくなる文書の組み合わせを抽出する文書抽出手段とを備えたことを特徴とする文書抽出装置。

【請求項 2】 上記類似度計算手段は、各文書を所定の文字列毎に分割する文字列分割機能部と、この文字列分割機能部で分割された文字列の出現頻度を基に各文書の文書ベクトルを算出する文字列頻度計算機能部と、この文字列頻度計算機能部で得られた文書ベクトルを基に各文書間の類似度を計算する相互類似度計算機能部とを備えたことを特徴とする請求項 1 に記載の文書抽出装置。

【請求項 3】 上記文字列分割機能部は、形態素解析、n - g r a m、ストップワードのいずれかの文字列分割方式を用いて上記各文書を所定の文字列毎に分割することを特徴とする請求項 2 に記載の文書抽出装置。

【請求項 4】 上記文字列頻度計算機能部は、分割された文字列の出現頻度を基に各文書を T F I D F で重み付けした文書ベクトルを生成することを特徴とする請求項 2 または 3 に記載の文書抽出装置。

【請求項 5】 上記相互類似度計算機能部は、各文書の文書ベクトルを基にベクトル空間法を用いて各文書間の類似度を計算することを特徴とする請求項 2 ～ 4 のいずれかに記載の文書抽出装置。

【請求項 6】 抽出候補となる複数の文書を取得し、それら各文書間の全ての類似度を算出する類似度計算手段と、その文書群のなかから任意の数の文書を抽出するに際し、この類似度計算手段で算出された各文書間の類似度の総和が最も小さくなる文書の組み合わせを抽出する文書抽出手段とをコンピュータに機能させることを特徴とする文書抽出プログラム。

【請求項 7】 上記類似度計算手段は、各文書を所定の文字列毎に分割する文字列分割機能と、この文字列分割部で分割された文字列の出現頻度を基に各文

書の文書ベクトルを算出する文字列頻度計算機能と、この文字列頻度計算部で得られた文書ベクトルを基に各文書間の類似度を計算する相互類似度計算機能とを備えたことを特徴とする請求項 6 に記載の文書抽出プログラム。

【請求項 8】 上記類似度計算手段は、形態素解析、n-gram、ストップワードのいずれかの文字列分割方式を用いて上記各文書を文字列毎に分割する文字列分割機能と、分割された文字列の出現頻度を基に T F I D F で重み付けした各文書の文書ベクトルを生成する文字列頻度計算機能と、各文書の文書ベクトルを基にベクトル空間法を用いて各文書間の類似度を計算する相互類似度計算機能と、を備えたことを特徴とする請求項 6 に記載の文書抽出プログラム。

【請求項 9】 抽出候補となる複数の文書を取得し、それら各文書間の全ての類似度を算出した後、その文書群のなかから任意の数の文書を抽出するに際して各文書間の類似度の総和が最も小さくなる文書の組み合わせを抽出するようにしたことを特徴とする文書抽出方法。

【請求項 10】 上記抽出候補となる各文書間の類似度は、各文書を所定の文字列毎に分割した後、分割された文字列の出現頻度を計算し、その文字列の出現頻度を基にして各文書の文書ベクトルを算出し、その文書ベクトルを用いて算出するようにしたことを特徴とする請求項 9 に記載の文書抽出方法。

【請求項 11】 上記抽出候補となる各文書間の類似度は、形態素解析、n-gram、ストップワードのいずれかの文字列分割方式を用いて各文書を所定の文字列毎に分割した後、分割された文字列の出現頻度を基に T F I D F で重み付けした各文書の文書ベクトルを算出し、その文書ベクトルを基にベクトル空間法によって算出するようにしたことを特徴とする請求項 9 に記載の文書抽出方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、ニュース等の文書をユーザの好みに応じて自動的に配信する文書配信システム等に係り、特に配信候補となる数多い文書の中から内容的に類似している文書を排除して特徴のある文書のみを抽出するための装置及びその抽出プロ

グラム並びに抽出方法に関するものである。

【0002】

【従来の技術】

ユーザ毎にカスタマイズが可能な情報配信システムは、ユーザがフィルタリング条件を設定し、リアルタイムで送られてくるニュース等の各種情報（以下、文字情報を主体とした文書という。）の中からコンピュータが自動的にその設定されたフィルタリング条件に合う文書のみを抽出してユーザに配信する形態が一般的である。

【0003】

このような形態の文書配信システムの場合、フィルタリング条件によっては、配信される文書が偏りすぎたり、また、同じような内容の文書が繰り返し送られてくるといった問題点がある。特に、後者の問題点に関しては、文書の内容が重複することにより、配信される情報に無駄が多くなったり、文書掲載スペースが限られている場合には他の重要な文書がカットされてしまう等の不都合を招き、文書配信システム自体の利便性や信頼性等を大きく損なう結果となる。

【0004】

そのため、このような文書の重複配信を防止すべく必要な文書のみを効率的に抽出するためのフィルタリング、あるいは分類技術が極めて重要となっており、これらに関する従来技術として、例えば以下の特許文献1及び2に示すような技術が提案されている。

先ず、特許文献1（特許第3203203号公報）には、すべての文書にキーワードを付与し、そのキーワードから文書をベクトル化し、ある文書Aが他の文書Bに包含されているときに最大値をとるような類似度評価尺度を導入して、代表文書、従属文書、独立文書等を認識して適宜関係のある文書をまとめる等の技術が開示されている。

【0005】

一方、特許文献2（特開2000-148770号公報）には、分類対象となる文書の特徴量を計算し、それら各特徴量の類似度を求めた後、数学的、統計的なクラスタ分析によって文書を分類する等の技術が開示されている。

【 0 0 0 6 】**【特許文献 1】**

特開平 9 - 9 6 4 1 8 号公報

【特許文献 2】

特開 2 0 0 0 - 1 4 8 7 7 0 号公報

【 0 0 0 7 】**【発明が解決しようとする課題】**

ところで、前者の従来技術にあっては、全ての文書にキーワード等の特徴を付与する必要があるが、全ての文書に対してキーワードを付与する作業には多くのコストがかかる。一方、後者の従来技術で用いているクラスタ分析は階層的な分類やグループ分けをするために適した分析方法であるが、文書数が増えると計算量が極端に多くなってしまい、スループットが大きく低下するといった問題点がある。

【 0 0 0 8 】

そこで、本発明は、このような従来の技術の有する未解決の課題に着目してなされたものであり、その目的は、低コストでかつ抽出に要する計算量も少ない新規な文書抽出装置及び文書抽出プログラム並びに文書抽出方法を提供するものである。

【 0 0 0 9 】**【課題を解決するための手段】**

上記課題を解決するために発明 1 の文書抽出装置は、抽出候補となる複数の文書を取得し、それら各文書間の全ての類似度を算出する類似度計算手段と、その文書群のなかから任意の数の文書を抽出するに際し、この類似度計算手段で算出された各文書間の類似度の総和が最も小さくなる文書の組み合わせを抽出する文書抽出手段とを備えたものである。

【 0 0 1 0 】

このような構成を採用することにより、抽出候補となる複数の文書のなかからいくつかの文書を抽出する際に、類似度が大きい文書同士は一緒に選択されることがなくなるため、内容が類似するような文書を重複して抽出する可能性を大幅

に低下させることができる。また、文書の抽出に際しては各文書にキーワードを付与する等の作業を要しないため、その作業に要するコストが不要となる。また、各文書間の類似度の総和によって文書の組み合わせを抽出するため、文書量が多くなった場合でも計算量が極端に増えてしまうようなこともない。

【0011】

また、発明2の文書抽出装置は、上記類似度計算手段を、各文書を所定の文字列毎に分割する文字列分割機能部と、この文字列分割機能部で分割された文字列の出現頻度を基に各文書の文書ベクトルを算出する文字列頻度計算機能部と、この文字列頻度計算機能部で得られた文書ベクトルを基に各文書間の類似度を計算する相互類似度計算機能部とで構成したものである。

【0012】

このような構成を採用することにより、各文書間の類似度を的確に算出することができ、発明1の効果を確実に実現することができる。

また、発明3の文書抽出装置は、上記文字列分割機能部が、形態素解析、n-gram、ストップワードのいずれかの文字列分割方式を用いて上記各文書を文字列毎に分割するようにしたものである。

【0013】

すなわち、これら形態素解析、n-gram、ストップワードといった文字列分割方式は従来から多用されている信頼性に優れた方式であり、これらを本発明の文字列分割機能部として用いることにより、各文書を的確に文字列に分割することは勿論、これらいずれかの方式を用いることにより様々な形態の文書にも的確に対応することができる。

【0014】

また、発明4の文書抽出装置は、上記文字列頻度計算機能部が、分割された文字列の出現頻度を基にTFIDFで重み付けした各文書の文書ベクトルを生成するようにしたものである。

すなわち、各文書の文書ベクトルを生成するに際し、分割された文字列の出現回数をそのまま用いても良いが、後述するTFIDFという文字列の重要度を反映した公知の重み付け方法を用いると各文書の特徴を良く表現した文書ベクトル

を生成できる。

【0015】

また、発明5の文書抽出装置は、上記相互類似度計算部が、各文書の文書ベクトルを基にベクトル空間法を用いて各文書間の類似度を計算するようにしたものである。

すなわち、各文書間の類似度を計算する方式としてベクトル空間法を用いれば、2つのベクトルの類似度は2つのベクトルのなす角の余弦（0～1）として定量的に表現することが可能となり、後の文書抽出をよりの確に行うことが可能となる。

【0016】

発明6の文書抽出プログラムは、抽出候補となる複数の文書を取得し、それら各文書間の全ての類似度を算出する類似度計算手段と、その文書群のなかから任意の数の文書を抽出するに際し、この類似度計算手段で算出された各文書間の類似度の総和が最も小さくなる文書の組み合わせを抽出する文書抽出手段とをコンピュータに機能させるようにしたものである。

【0017】

これにより、上述した各手段を実施するに際し、専用のハードウェアを用意することなくソフトウェアによって安価な汎用のパソコンをそのまま使用することができ、実施等に要するコストの大幅な削減や実施までの準備期間等を大幅に短縮することが可能となる。

また、発明7の文書抽出プログラムは、上記類似度計算手段を、各文書を所定の文字列毎に分割する文字列分割機能と、この文字列分割部で分割された文字列の出現頻度を基に各文書の文書ベクトルを算出する文字列頻度計算機能と、この文字列頻度計算部で得られた文書ベクトルを基に各文書間の類似度を計算する相互類似度計算機能とで実現したものである。

【0018】

これにより、発明2と同様に各文書間の類似度をソフトウェアによりの確に算出することができる。

また、発明8の文書抽出プログラムは、上記類似度計算手段を、形態素解析、

n-gram、ストップワードのいずれかの文字列分割方式を用いて上記各文書を文字列毎に分割する文字列分割機能と、分割された文字列の出現頻度を基にTFIDFを用いて各文書の文書ベクトルを生成する文字列頻度計算機能と、各文書の文書ベクトルを基にベクトル空間法を用いて各文書間の類似度を計算する相互類似度計算機能とで実現したものである。

【0019】

これにより、発明3～5と同様な作用・効果をソフトウェア上で確実に達成することができる。

発明9の文書抽出方法は、抽出候補となる複数の文書を取得し、それら各文書間の全ての類似度を算出した後、その文書群のなかから任意の数の文書を抽出するに際して各文書間の類似度の総和が最も小さくなる文書の組み合わせを抽出するようにしたものである。

【0020】

これによって、発明1の文書抽出装置と同様に、内容が類似（重複）するような文書を抽出する可能性を大幅に低下させることができると共に、文書抽出処理に要するコストが安価となり、文書量が多くなった場合でも計算量が大幅に増えてしまうようなこともない。

また、発明10の文書抽出方法は、上記抽出候補となる各文書間の類似度は、各文書を所定の文字列毎に分割した後、分割された文字列の出現頻度を計算し、その文字列の出現頻度を基にして各文書の文書ベクトルを算出し、その文書ベクトルを用いて算出するようにしたものである。

【0021】

これにより、発明2と同様に、各文書間の類似度を的確に算出することができる。

また、発明11の文書抽出方法は、上記抽出候補となる各文書間の類似度は、形態素解析、n-gram、ストップワードのいずれかの文字列分割方式を用いて各文書を所定の文字列毎に分割した後、分割された文字列の出現頻度を基にTFIDFを用いて各文書の文書ベクトルを算出し、その文書ベクトルを基にベクトル空間法によって算出するようにしたものである。

【0022】

これにより、発明 3～5 と同様な作用・効果を確実に達成することができる。

【0023】**【発明の実施の形態】**

以下、本発明の実施の形態を添付図面を参照しながら詳述する。

先ず、図 1 は本発明に係る文書抽出装置 10 の実施の一形態を示したものである。図示するように、この文書抽出装置 10 はインターネット等の情報通信網内にある情報供給源 S から供給されるいくつかの情報をそれぞれ文書として一時的に記憶しておく情報記憶手段 12 と、この情報記憶手段 12 に記憶された複数の文書をまとめて取得してそれら各文書間の類似度を算出する類似度計算手段 14 と、この類似度計算手段 14 で得られた各文書間の類似度を基にしてその文書群の中からいくつかの文書のみを抽出する文書抽出手段 16 とから主に構成されている。

【0024】

また、図示するようにこの類似度計算手段 14 はさらに、文字列分割機能部 18 と、文字列頻度計算機能部 20 と、相互類似度計算機能部 22 とから構成されており、後に詳述するが、情報記憶手段 12 から取得された各文書に対し、文字列分割機能部 18 によって各文書を文字列毎に分割した後、分割された各文字列の出現頻度を文字列頻度計算機能部 20 によって算出して各文書のベクトルを算出し、その後、この文字列頻度計算機能部 20 で得られた各文書の文書ベクトル相互の類似度を相互類似度計算機能部 22 によって算出してそのデータを求めるようになっている。

【0025】

具体的にはこの文書抽出装置 10 は図 2 に示すような構成をしたコンピュータ 100 により実現されることになる。

図示するように、このコンピュータ 100 は、制御プログラムに基づいて演算および装置全体を制御する CPU 30 と、所定領域にあらかじめ CPU 30 の制御プログラム等を格納している ROM 32 と、ROM 32 等から読み出したデータや CPU 30 の演算過程で必要な演算結果を格納するための RAM 34 と、外

部装置に対してデータの入出力を媒介する I / F 3 8 とで構成されており、これらは、データを転送するための信号線であるバス 3 9 で相互にかつデータ授受可能に接続されている。

【 0 0 2 6 】

I / F 3 8 には、外部装置として、データ入力可能なキーボードやマウス等からなる入力装置 4 0 と、画像信号に基づいて画面を表示する表示装置 4 2 と、前述したように情報供給源 S から供給される情報を所定の文書データとして一時的に記憶するための情報記憶手段 1 2 とが接続されている。この情報記憶手段 1 2 は、例えばハードディスク等の外部記憶装置であり、インターネット等の情報供給源 S から所定の情報が定期的にあるいは随時供給されるようになっている。

【 0 0 2 7 】

そして、CPU 3 0 は、マイクロプロセッシングユニット MPU 等からなり、ROM 3 2 の所定領域に格納されている文書抽出プログラムを起動させ、その文書抽出プログラムに従って上記類似度計算手段 1 4 に相当する処理と上記文書抽出手段 1 6 に相当する処理をそれぞれ時分割で実行するようになっている。

以上において、本実施の形態の動作を説明する。

【 0 0 2 8 】

図 1 に示すように、先ず情報記憶手段 1 2 には情報供給源 S からユーザの好みに対応した文書が一定の時間毎もしくは不定期に供給されて一時的に蓄積され、その文書数が所定数に達したとき、あるいは保存時間が一定時間経過したならば、一旦、その蓄積された文書の全てが類似度計算手段 1 4 に送られ、ここで各文書間の類似度が算出される。

【 0 0 2 9 】

すなわち、類似度計算手段 1 4 に送られてきた各文書は、先ず文字列分割機能部 1 8 によって文字列に分割される。この文字列の分割方式（手法）は特に限定されるものではないが、図 3 に示すように各文書 $D_1 \sim D_m$ を文字列に分割するに際して形態素解析を用いた場合は、形態素解析辞書を参照しながら文法的な区切りで文字列（単語）に分割することができる。ここで、この形態素解析には様々な手法があり、辞書の善し悪しによっても結果は異なるが、例えば同図の「無

線／の／セキュリティ／が／話題／に／なっ／ている／。／…」等のように、名詞、動詞、形容詞助詞、助動詞等の単語に分けることができる。また、この形態素解析は分割の精度が良いが、以前では精度を維持するために辞書の作成やメンテナンスにコストは掛かるといった欠点があったが、最近では長年十分に作り込まれてきた辞書が資産として使えるため、コストの問題も次第に解消されてきており、現在最もよく使われる文字列分割方法である。ただし、この形態素解析は日本語に限って使用できるものであり、英語や中国語などの他の言語には使えないといった不利な面もある。

【0030】

また、このように各文書 $D_1 \sim D_m$ を文字列に分割するに際して形態素解析ではなく、一定間隔毎に文字列を切っていく $n\text{-gram}$ という文字列分割方式を用いることも可能であり、この $n\text{-gram}$ 方式を用いた場合、上記文書は図4に示すように分割される。すなわち、この $n\text{-gram}$ の「 n 」とは何バイト毎（または何文字毎）かを表す数字で、図4の場合では2文字毎なので 2-gram と書くことができる。ただし、日本語などの2バイト文字の場合、2文字＝4バイトなので 4-gram と書く場合もあるかもしれないが、ここではその数字の正確さは問題にすることではない。 $n\text{-gram}$ は、意味のある単語を塊として切り出すことは困難であるが、分割したものをそのまま統計的に処理するだけであれば必ずしも意味のある単語が塊になっている必要がない場合もある。また、この $n\text{-gram}$ は上記形態素解析に比べてアルゴリズムが単純でどの言語に対しても使えるというメリットがある。

【0031】

また、他の文字列分割方式として図5に示すようなストップワードという方式を使用しても良い。このストップワード方式とは文書の中で切れ目となる文字や規則を登録し、それに従って分割していく方法である。例えば、図5に示す例では、①助詞だと思われる「の」「は」「が」「に」「を」「や」、②句読点「、」「。」、③漢字、カタカナ、アルファベット等の字種の変わり目、等といった3つのルールのいずれかが成立するところで分割したものである。尚、このストップワードはある程度意味のある単語を抜き出すことが可能であるが、「情報通

信技術」等といった長い熟語や「インターネットテクノロジー」等といった長いカタカナの複合語などは分割できないという問題もある。また、英語であれば、①スペース、②カンマ、ピリオド、コロンの、セミコロン、その他の記号、③アルファベット、数字、記号などの字種の変わり目等といったルールをもとに、単語の活用形を落とすステミングという手法を使うことである程度の文字列分割を行うことができる。

【0032】

このように上記文字列分割機能部18によって全ての文書 $D_1 \sim D_m$ についての文字列分割が行われたならば、次に文字列頻度計算機能部20によって文字列頻度の計算を行い、図7に示すような文字列－文書行列を作成する。この文字列－文書行列は各文書 $D_1 \sim D_m$ とユニークな文字列 $T_1 \sim T_n$ との対応関係を示したものであり、各文字列 $T_1 \sim T_n$ が各文書 $D_1 \sim D_m$ 中に何回出現するかを数え、それを示したものである。例えば、文字列分割方式として形態素解析を用いた分割結果の場合では、図6に示すように、文書 D_1 中には「無線」（網掛け文字）という文字列（ T_1 ）は3回出現しており、その W_{11} に相当する行列の要素は、その出現回数をそのまま用いた場合では「3」となる。

【0033】

ここで、 W_{mn} に相当する行列の各要素は、文字列の出現回数をそのまま用いても良いが、TFIDF（Term Frequency & Inverse Document Frequency）という文字列の重要度を反映した重み付け方法を用いると、各文書の特徴を良く表現した文書ベクトルが生成できることが知られており、後の相互類似度計算で活用することができる。

【0034】

すなわち、このTFIDFは、以下の数式1に示すように、ある文書D内での文字列Tの出現頻度（TF：Term Frequency）と、文書集合全体で文字列Tが出現する文書数の頻度を逆数（IDF：Inverse Document Frequency）の積で求め、数値が大きいほどその文字列Tが重要であることを表すものである。TFは頻出する文字列は重要であるという指標であり、ある文書中に文字列が出現する頻度が増加すると大きくなる性質を持っている。IDFは多くの文書中に出現する

文字列は重要でない、つまり特定の文書に出現する文字列が重要であるという指標であり、ある文字列が使われている文書数が減少すると大きくなるという性質を持っている。従って、このTFIDFの値は、頻繁に出現するが多くの文書に出現する文字列（接続詞や助詞等）や、特定の文書にのみ出現するがその文書に高頻度で出現する文字列に対しては大きくなる性質を持っており、このTFIDFによって文書中の文字列は数値化され、その数値を要素として文書をベクトル化することができる。

【0035】

【数1】

$$W(t, d) = TF(t, d) \times IDF(t)$$

但し、 $TF(t, d)$ = 文書dに文字列tが出現する頻度

$$IDF(t) = \log\left(\frac{D}{DF(t)}\right)$$

$DF(t)$ = 文書全体で文字列tが出現する文書数の頻度
 D = 全文書数

【0036】

このような文書取得から文字列計算までの流れを表したのが図8に示すフローチャートである。図示するように、ステップS100では情報記憶手段12にストックされた文書の一つずつ取得し、ステップS102では取得したその文書を文字列毎に分割し、ステップS104ではその文書と文字列との対応関係を示す文字列－文書行列に文字列の頻度情報を記憶してステップS106に移行する。ステップS106では情報記憶手段12にストックされている文書が残っているか否かを判定し、残っている場合（YES）には、その文書を取得して同様な処理を行い、全ての文書がなくなるまで繰り返す。

【0037】

一方、ステップS106において情報記憶手段12にストックされている文書が残っていないと判定した場合（NO）にはステップS108に移行し、完成した文字列－文書行列の頻度情報を基にTFIDFによって重み付けし直した文字列－文書行列を作成する。これによって全ての文書はそれらに出現するユニーク

な文字列の数と同じ次元（数千～数十万）のベクトルとして表現できることになる。

【0038】

そして、このようにして全ての文書がベクトル化されたならば、相互類似度計算機能部22によって各文書間の類似度が求められる。具体的には、この相互類似度計算機能部22は公知のベクトル空間法を採用するものであり、上記TFIDFによって求められた各文書ベクトルは、このベクトル空間法によって相互の類似度が定義されることになる。すなわち、対比する2つの文書のベクトル類似度は、図9に示すように2つのベクトルのなす角 θ の余弦値（0～1）として定義できることから、各文書同士の類似度は図10に示すような対称行列で表現できる。

【0039】

その後、その対称行列をもとに類似情報のグルーピングやカットを行うことで類似文書を排除したフィルタリングが実現可能となる。例えば、図10のような対称行列では文書 D_1 と文書 D_2 の類似度は図11に示すように0.9、文書 D_1 と文書 D_3 の類似度は0.3というように各文書間の類似度が定量的に示される。

【0040】

次に、このように類似度計算手段14によって各文書間の類似度が定量的に求められたならば、文書抽出手段16が、その文書群のなかから各文書 $D_1 \sim D_n$ 間の類似度の総和が最も小さくなる文書の組み合わせを抽出することになる。

具体的に説明すると、この文書抽出手段16は、取得された全ての文書 n 個の中から r 個（文書の配信量やレイアウトの都合等によって決定される。）の文書を抽出することになるが、この r 個の文書を抽出するに際しては、以下の数式2に示すように n 個の文書の中から r 個を選ぶ全ての文書の組み合わせを検討し、それら各組み合わせの類似度を合計し、その総和が最も小さいものを抽出することになる。

【0041】

【数 2】

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

【0042】

例えば、抽出候補となる文書が4つ (D_1, D_2, D_3, D_4) あり、その中から3つの文書を抽出する組み合わせは以下の数式3に示すように4通りとなる。

【0043】

【数 3】

$${}_4C_3 = \frac{4!}{3!(4-3)!} = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (1)} = 4$$

【0044】

そして、それぞれの組み合わせからできる2つ組の類似度を図11に示す類似度対称行列からピックアップしていき、以下の表1に示すように各組み合わせI、II、III、IV毎にそれらの類似度を単純に足してその総和を計算する。

【0045】

【表 1】

I	$D_1, D_2, D_3 \Rightarrow (D_1, D_2)=0.9, (D_1, D_3)=0.3, (D_2, D_3)=0.2 \Rightarrow 0.9+0.3+0.2=1.4$
II	$D_1, D_2, D_4 \Rightarrow (D_1, D_2)=0.9, (D_1, D_4)=0.5, (D_2, D_4)=0.8 \Rightarrow 0.9+0.5+0.8=2.2$
III	$D_1, D_3, D_4 \Rightarrow (D_1, D_3)=0.3, (D_1, D_4)=0.5, (D_3, D_4)=0.3 \Rightarrow 0.3+0.5+0.3=1.1$
IV	$D_2, D_3, D_4 \Rightarrow (D_2, D_3)=0.2, (D_2, D_4)=0.8, (D_3, D_4)=0.3 \Rightarrow 0.2+0.8+0.3=1.3$

【0046】

表1に示す例では、4つの組み合わせI、II、III、IVのうちII (D_1, D_2, D_4) の組み合わせの類似度の総和が2.2と最も高く、III (D_1, D_3, D_4) の組み合わせの類似度の総和が1.1と最も低くなったことから、4つの文書 (D_1, D_2, D_3, D_4) のなかから文書 D_1, D_3, D_4 の文書を組み合わせで抽出することになる。

【0047】

このように本発明は各文書間の類似度の総和が最も小さくなる文書の組み合わ

せを抽出するようにしたことから、各文書の類似度が小さくなり、内容が似通った文書を同時に抽出する可能性を大幅に減少することができる。

この結果、これを前述した文書配信システム等に適用した場合、重複した内容の文書がユーザに配信される等といった不都合を未然に回避することが可能となると共に、抽出に際して文書毎にキーワードを付与するなどといった煩わしい作業が不要となり、文書抽出処理に要するコストの削減に大きく貢献することができる。また、各文書間の類似度の算出に際しては計算量が大きくなる可能性のあるクラスタ分析を用いる必要がなくなり、非力なコンピュータ等によってもその機能を十分発揮することが可能となる。

【0048】

尚、本実施の形態では類似度の総和として、定量的に求められた各文書間の類似度を単に足した数値を用いたが、その他に二乗和、対数和等を用いても良い。ただし、本実施の形態のように文書ベクトルの類似度は余弦を使って0～1の間の値として正規化されているので、敢えて二乗和や対数和等の非線形な関数を使う必要はなく、そのまま足し算することで十分である。

【図面の簡単な説明】

【図1】 文書抽出装置の構成を示すブロック図である。

【図2】 コンピュータの構成を示すブロック図である。

【図3】 形態素解析による文字列分割の一例を示す図である。

【図4】 n-gramによる文字列分割の一例を示す図である。

【図5】 ストップワードによる文字列分割の一例を示す図である。

【図6】 形態素解析による文字列分割結果を示す図である。

【図7】 文字列－文書行列を示す図である。

【図8】 文字列－文書行列を得るための流れを示すフローチャート図である。

【図9】 文書ベクトル及びその相関関係を示す図である。

【図10】 文書－文書間の対称行列を示す図である。

【図11】 文書－文書間の対称行列を示す図である。

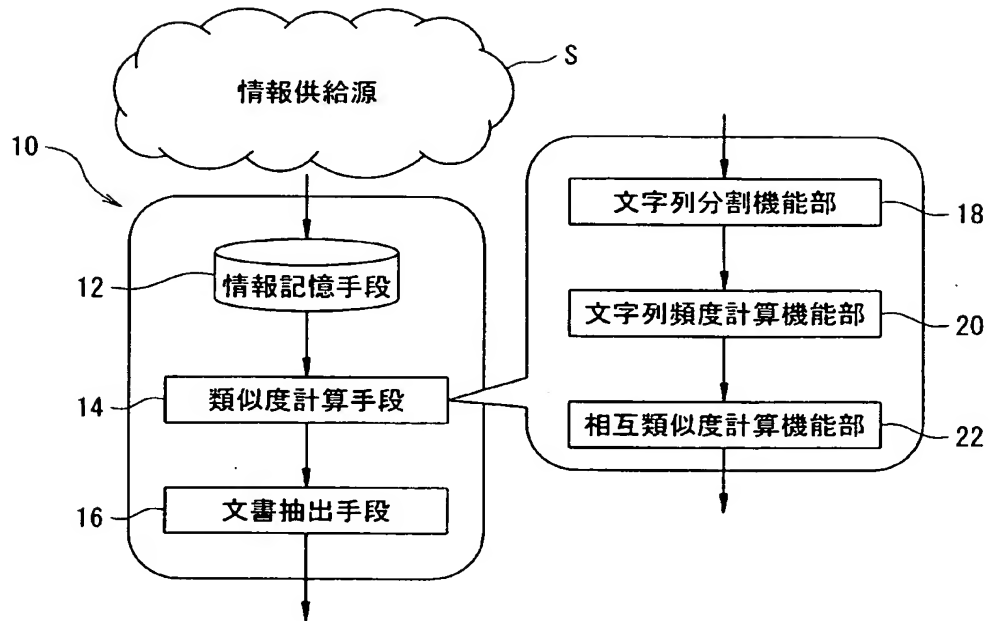
【符号の説明】

10…文書抽出装置、12…情報記憶手段、14…類似度計算手段、16…文書抽出手

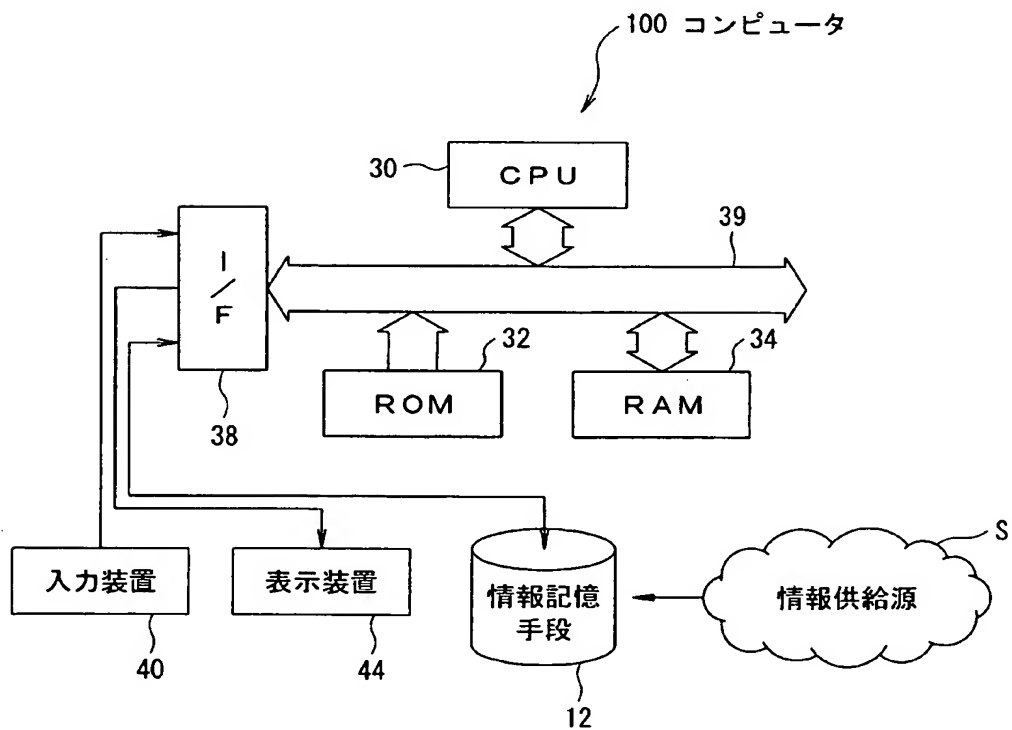
段、18…文字列分割機能部、20…文字列頻度計算機能部、22…相互類似度計算機能部、100…コンピュータ、 $D_1 \sim D_m$ …文書、 S …情報供給源

【書類名】 図面

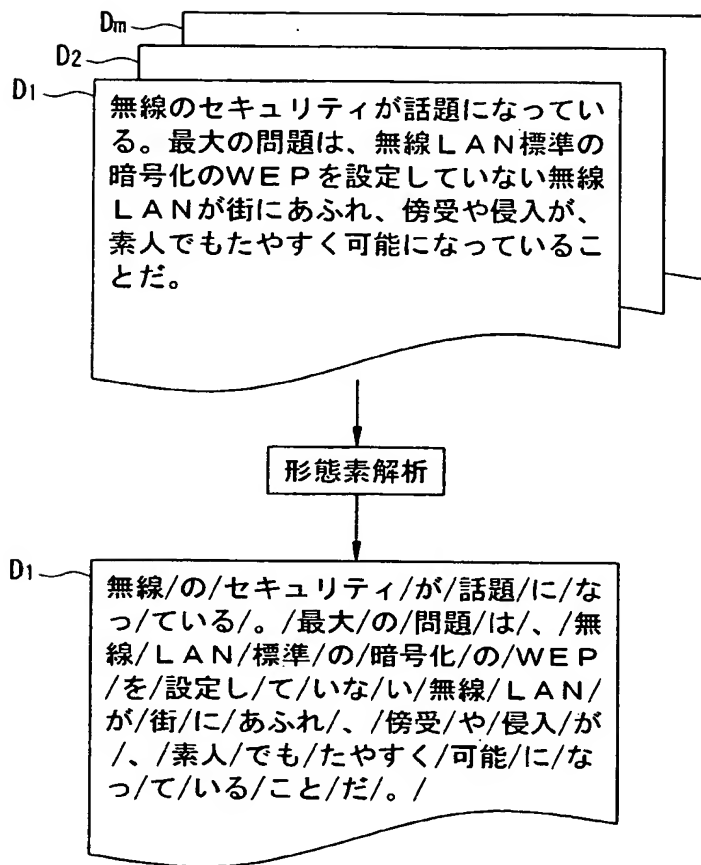
【図 1】



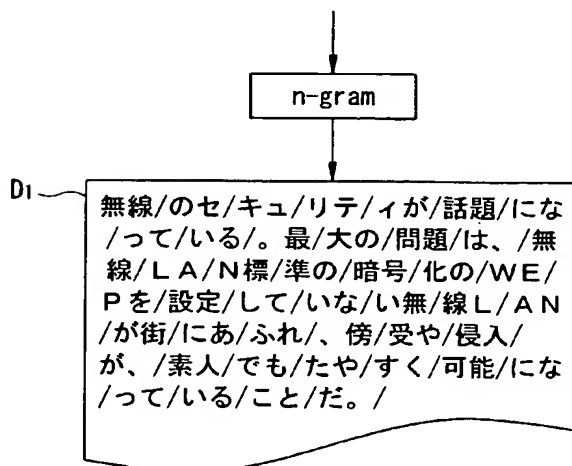
【図 2】



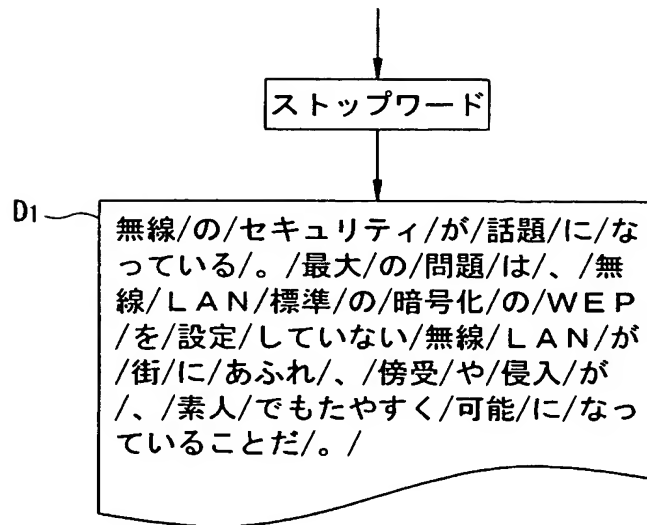
【図 3】



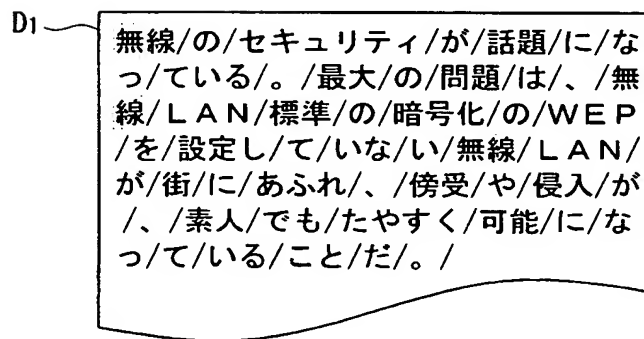
【図 4】



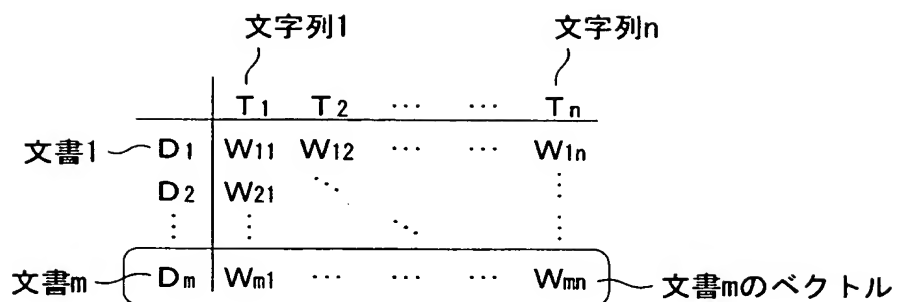
【図 5】



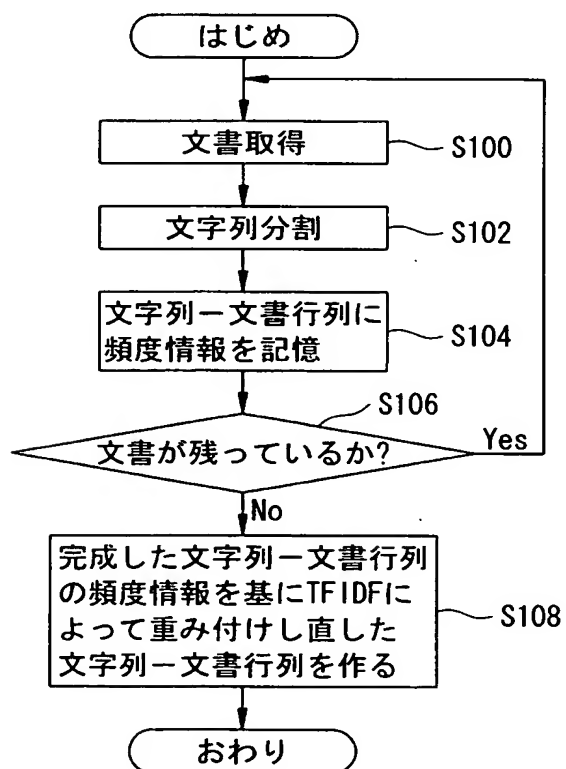
【図 6】



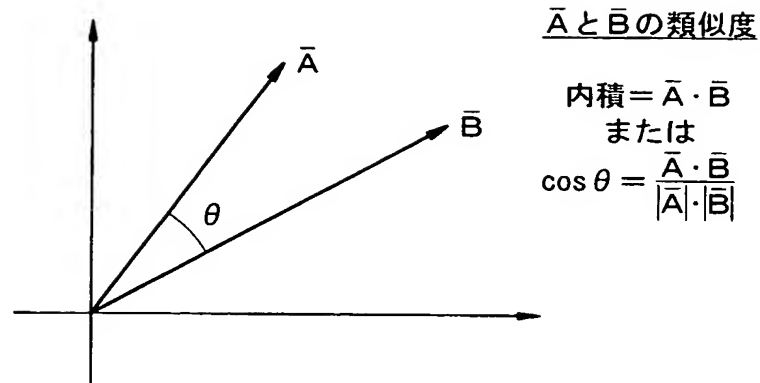
【図 7】



【図 8】



【図 9】



【図 10】

	D ₁	D ₂	D _m
D ₁		S ₁₂	S _{1m}
D ₂	S ₁₂				⋮
⋮	⋮				⋮
D _m	S _{1m}				

【図 11】

	D ₁	D ₂	D ₃	D ₄
D ₁		0.9	0.3	0.5
D ₂			0.2	0.8
D ₃				0.3
D ₄				

【書類名】 要約書

【要約】

【課題】 低コストでかつ抽出に要する計算量も少ない新規な文書抽出装置及び文書抽出プログラム並びに文書抽出方法の提供。

【解決手段】 抽出候補となる複数の文書 $D_1 \sim D_m$ 間の全ての類似度を算出する類似度計算手段14と、その文書群のなかから任意の数の文書 D を抽出するに際し、この類似度計算手段14で算出された各文書 D 間の類似度の総和が最も小さくなる文書の組み合わせを抽出する文書抽出手段16とを備える。これにより、文書 D の抽出に際しては各文書 $D_1 \sim D_m$ にキーワードを付与する等の作業に要するコストが不要となり、また、文書量が多くなった場合でも抽出に要する計算量が大幅に増えてしまうようなこともない。

【選択図】 図1

認定・付加情報

特許出願の番号	特願 2 0 0 2 - 3 6 0 9 8 4
受付番号	5 0 2 0 1 8 8 4 5 1 2
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 4 年 1 2 月 1 3 日

< 認定情報・付加情報 >

【提出日】 平成14年12月12日

次頁無

特願 2 0 0 2 - 3 6 0 9 8 4

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 2 3 6 9]

1 . 変更年月日

1 9 9 0 年 8 月 2 0 日

[変更理由]

新規登録

住 所

東京都新宿区西新宿 2 丁目 4 番 1 号

氏 名

セイコーエプソン株式会社